DOCUMENT RESUME

ED 420 695 TM 028 370

AUTHOR Nering, Michael L.
TITLE The Influence of Nonmodel-Fitting Examinees in Estimating Person Parameters.
PUB DATE 1998-04-13
NOTE 31p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Ability; *Estimation (Mathematics); *Item Response Theory; Models
IDENTIFIERS *Calibration; *Person Fit Measures

ABSTRACT

The issue of person fit has received an increasing amount of attention by researchers in the past few years. Several studies have focused on the issue of how nonmodel-fitting responses affect the accuracy of ability estimates (e.g. R. Meijer and S. Nering, in press; Reise, 1995). The purpose of this study was to examine the effects that nonmodel-fitting response vectors (NRVs) have on the estimation of person parameters for model-fitting response vectors (MRVs). Under the assumption of local dependence in item response theory, one examinee should not influence the test results of other examinees. However, if NRVs are present in a calibration sample, and they affect the quality of item parameter estimates, this could cause error in person parameter estimates for MRVs. In this study, ability estimates and person-fit statistics were estimated for MRVs in calibration samples that contained different amounts and types of NRVs. It was found that an expected a posteriori estimation procedure tended to result in reduced bias in ability estimates, while a biweight estimation procedure resulted in person-fit statistics that were more normally distributed. The ZU3 nonparametric person-fit index was much less sensitive to NRVs compared to the parametric person-fit statistic. The results of this study demonstrated that researchers should consider the calibration sample seriously when estimating person parameters. (Contains 5 figures, 4 tables, and 29 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 420 695

Running Head: ESTIMATING PERSON PARAMETERS

The Influence of Nonmodel-Fitting Examinees in Estimating Person Parameters

Michael L. Nering

ACT, *Inc.*

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Michael Nering

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM028370

Abstract

The issue of person fit has received an increasing amount of attention by researchers in the past few years. Several studies have focused on the issue of how nonmodel-fitting responses effect the accuracy of ability estimates (e.g., Meijer & Nering, in press; Reise, 1995). The purpose of this study was to examine the effects that nonmodel-fitting response vectors (NRVs) have on the estimation of person parameters for model-fitting response vectors (MRVs). Under the assumption of local dependenc in item response theory, one examinee should not influence the test results of other examinees. However, if NRVs are present in a calibration sample, and they affect the quality of item parameter estimates, and this could cause error in person parameter estimates for MRVs. In this study ability estimates and person-fit statistics were estimated for MRVs in calibration samples that contained different amounts and types of NRVs. It was found that an expected a posterior estimation procedure tended to result in reduced bias in ability estimates, while a Biweight estimation procedure resulted in person-fit statistics that were more normally distributed. The ZU3 nonparametric person-fit index was much less sensitive to NRVs compared to the parametric person-fit statistic l_z. The results of this study demonstrated that researchers should seriously consider the calibration sample when estimating person parameters.

Key words: Person fit, appropriateness measurement, ability estimation, nonparametric item response theory, item response theory.

The Influence of Nonmodel-Fitting Examinees in Estimating Person Parameters

The extent to which an ability estimate $(\hat{\theta}_j)$ for an examinee j represents the underlying latent trait of interest (θ) is commonly referred to as person fit (e.g., Nering, 1995). This area of research has recently received a great deal of attention (e.g., Meijer & Nering, in press; Meijer, 1996; Meijer & Sijtsma, 1995), and has continued to assist researchers in studying a variety of measurement related problems (e.g., Reise & Waller, 1993; Schmitt, Cortina, & Whitney, 1993; Zickar & Drasgow, 1996). Although this area of research has continued to expand over the past few decades, many issues regarding person fit continue to be problematic for measurement specialists. For example, as discussed by Nering and Meijer (in press) most methods used to index person fit cannot be used to determine the cause of the nonmodel-fitting behavior. Methods that can be used to determine why an examinee may have a poorly estimated θ may be useful in selecting candidates for a job, in a college admissions test, or within the context of classroom assessment.

As discussed by Levine and Rubin (1979) and Wright (1977), there are many examinee response behaviors that could potentially lead to a poorly estimated θ value. For example, an examinee may lack English skills necessary to answer a test question used in measuring mathematics ability, or an he or she may be overly anxious or experience fatigue on longer tests. Nering (1996) suggests that examinees in a computerized adaptive test may lack computer familiarity or may experience a warm-up effect, which could result in a poorly estimated θ value. Behaviors such as these may result in an examinee responding to test questions in a manner that does not reflect his or her θ, that is, in a manner that is not in accordance with the underlying test model. Certainly, as researchers continue to find new methods for indexing person fit and new applications for using person fit, this topic will play an important role in educational and psychological measurement.

Although an increasing number of researchers have found applications for which person fit is useful, much of the work done in this area has focused on the development of person-fit statistics, on the detection of person fit, or on the influence that nonmodel-fitting responses has on ability estimation. The purpose of the present investigation, however, was to study the influence that nonmodel-fitting response

vectors (NRVs) have on the estimation of person parameters for model-fitting response vectors (MRVs). Past researchers have typically only focused on how accurately $\hat{\theta}$ values are estimated for NRVs [i.e., manipulated response vectors (e.g., Meijer, 1996; Nering, 1996)]. Ideally, within the context of item response theory (IRT), one examinee should not influence another examinee's $\hat{\theta}$; however, in a real testing situation an examinee's $\hat{\theta}$ is determined using estimated item parameters. Estimated item parameters are typically found from a calibration sample of examinees using a program such as BILOG (Mislevy & Bock, 1982). If the sample contains NRVs then the item parameters may be poorly estimated, and these poorly estimated item parameters will affect $\hat{\theta}$ values for MRVs, and one goal of this study was to determine the extent to which these $\hat{\theta}$ values are affected. Along with the estimation of θ, NRVs may influence other estimated person parameters, such as indices of person fit. Thus, a second goal of this study was to investigate how NRVs affect indices of person fit for MRVs.

A schematic of the basic research questions studied here can be best summarized by Figure 1. Notice that there are two datasets, which are similar except that the dataset to the right contains NRVs. The comparisons made in this study were in the estimated person parameters for the MRVs under different calibration conditions.

Past Person Fit Research

In the past 20 years several methods have been developed to index person-model fit, and to detect examinees whose $\hat{\theta}$ may not be accurately estimated. Most of the person-fit indexing methods developed within the context of parametric IRT models can be categorized into two general approaches (Nering, 1997). One approach uses the peak of the likelihood function of a response pattern (Drasgow, Levine, & Williams, 1985; Levine & Rubin, 1979), and the other approach evaluates the discrepancy between observed and model predicted responses (Tatsuoka, 1984; Trabin & Weiss, 1983). Many studies have been conducted that have compared how well different person-fit indices detect NRVs under a variety of conditions (e.g., Birenbaum, 1985, 1986; Drasgow, Levine, & McLaughlin, 1987, 1991).

Although a single person-fit index has not been found to be uniformly superior at detecting NRVs, the l_z index has received a great deal of attention. For example, researchers have studied the

distributional characteristics of l_z (Nering, 1995, 1997) and the detection rate of l_z (Reise & Due, 1991; Nering 1996) under many experimental conditions. These studies, along with others (e.g., Li & Olejnik, 1997), have shown that the l_z index may be the most promising tool in detecting NRVs within the context of the two and three-parameter logistic IRT models. This index can be defined as:

$$l_z = \frac{l_o - \mathrm{E}(l_o)}{\left[\mathrm{var}(l_o)\right]^{1/2}} \tag{1}$$

where, l_o represents the log of the peak of a likelihood function for a particular response pattern and $\mathrm{E}(l_o)$ and $\mathrm{var}(l_o)$ represent the expected value and the variance of l_o, respectively. The terms in Equation 1 can be computed by:

$$l_o = \ln \prod_{i=1}^{n} P_i(\hat{\theta})^{u_i} Q_i(\hat{\theta})^{1-u_i}, \tag{2}$$

$$E(l_o) = \sum_{i=1}^{n} \left[P_i(\hat{\theta}) \ln P_i(\hat{\theta}) + Q_i(\hat{\theta}) \ln Q_i(\hat{\theta}) \right], \text{ and} \tag{3}$$

$$\mathrm{var}(l_o) = \sum_{i=1}^{n} P_i(\hat{\theta}) Q_i(\hat{\theta}) \left\{ \ln \left[\frac{P_i(\hat{\theta})}{Q_i(\hat{\theta})} \right] \right\}^2 \tag{4}$$

respectively, where

i indexes the items (i = 1, 2, ... n),
P is the probability of a correct response given the IRT model,
Q is 1-P, and
u represents the item score (0 or 1).

Although the distribution of l_z has been studied under many conditions by past researchers, how NRVs affect the distribution of person fit for MRVs has not been investigated. If NRVs degrade how well

item parameters are estimated, and how well θ values are estimated for MRVs, then it is possible that NRVs will also affect how well other examinees appear to be fitting the underlying test model.

Indexing the extent to which $\hat{\theta}$ is an accurate representation of the latent trait of interest is not limited to a context in which parametric IRT models are used. For example, the ZU3 index can be used within the context of nonparametric IRT (Mokken, 1971; Mokken & Lewis, 1983). This ZU3 index was developed by van der Flier (1982) and can be defined by:

$$ZU3 = \frac{U3 - E(U3)}{V(U3)^{1/2}} \tag{5}$$

If we let g represent an item index ($g = 1, 2, \ldots, k$), r represent the total score for an examinee, π_g represent the difficulty of item g (i.e., proportion correct score), and X_g represent the item score (0/1) then the terms in Equation 5 can be found by:

$$U3 = \frac{\sum_{g=1}^{r} \ln\left(\frac{\pi_g}{1-\pi_g}\right) - \sum_{g=1}^{r} X_g \ln\left(\frac{\pi_g}{1-\pi_g}\right)}{\sum_{g=1}^{r} \ln\left(\frac{\pi_g}{1-\pi_g}\right) - \sum_{g=k-r+1}^{k} \ln\left(\frac{\pi_g}{1-\pi_g}\right)}, \tag{6}$$

$$E(U3) = \frac{\sum_{g=1}^{r} \ln\left(\frac{\pi_g}{1-\pi_g}\right) - \eta}{\sum_{g=1}^{r} \ln\left(\frac{\pi_g}{1-\pi_g}\right) - \sum_{g=k-r+1}^{k} \ln\left(\frac{\pi_g}{1-\pi_g}\right)}, \text{ and} \tag{7}$$

$$V(U3) = \frac{\beta}{\left[\sum_{g=1}^{r} \ln\left(\frac{\pi_g}{1-\pi_g}\right) - \sum_{g=k-r+1}^{k} \ln\left(\frac{\pi_g}{1-\pi_g}\right)\right]^2}, \text{ where} \tag{8}$$

$$\eta = \sum_{g=1}^{k} \pi_g \ln\left(\frac{\pi_g}{1-\pi_g}\right) + \frac{\sum_{g=1}^{k} \pi_g(1-\pi_g)\ln\left(\frac{\pi_g}{1-\pi_g}\right)}{\sum_{g=1}^{k} \pi_g(1-\pi_g)}\left(r - \sum_{g=1}^{k} \pi_g\right), \text{ and} \tag{9}$$

$$\beta = \sum_{g=1}^{k} \pi_g(1-\pi_g)\left[\ln\left(\frac{\pi_g}{1-\pi_g}\right)\right]^2 - \frac{\left[\sum_{g=1}^{k} \pi_g(1-\pi_g)\ln\left(\frac{\pi_g}{1-\pi_g}\right)\right]^2}{\sum_{g=1}^{k} \pi_g(1-\pi_g)}. \tag{10}$$

Researchers have shown that ZU3 has high detection rates under many conditions (Meijer, 1996); however, little work has been done concerning the distributional characteristics of this index. Like the l_z index, the ZU3 index standardized and should have an expected value of 0.0 and a standard deviation of 1.0. Additional work is needed to determine if this ZU3 index is distributed as expected, and to determine if NRVs affect the distribution of ZU3 for MRVs.

Ability Estimation

Person-fit researchers have not only been interested in the detection of lack of model fit, but also how different estimation procedures perform in the presense of NRVs. Reise (1995) found that Biweight estimation (BIW; Mislevy & Bock, 1982) resulted in higher detection rates for l_z compared to maximum likelihood estimation (MLE) and expected a posterior estimation (EAP). Although the detection rates using BIW were only slightly better, Reise's findings do suggest that the manner in which $\hat{\theta}$ is estimated may alleviate the effects of some types of nonmodel-fitting response behaviors.

The BIW procedure is a robust estimation method in that it down-weights items that have difficulty values that are different from an examinee's ability. As discussed by Reise and Due (1991), for an examinee to be considered nonmodel fitting, it is necessary that they respond inappropriately to items that they (probabilistically) should have gotten right or wrong. Thus, by down weighting the items where

there is a large difference in item difficulty and the examinee's ability. BIW may result in $\hat{\theta}$ values that are a more accurate representation of θ when nonmodel-fitting responses are present.

Meijer and Nering (in press) were able to extend Reise's (1995) findings and show that not only were there better detection rates when BIW was used, but also that BIW resulted in less bias in $\hat{\theta}$ for NRVs compared to when MLE and EAP were used. The findings in Meijer and Nering, however, were specific to the tails of the θ distribution, and examinees at or around θ=0.0 were less affected by what estimation method was used.

Both the Reise (1995) and the Meijer and Nering (in press) studies only investigated the influence that nonmodel-fitting responses had on the ability estimation for NRVs, and additional work is needed to determine what influence these response behaviors have on MRVs. Moreover, it is necessary to determine if there is an estimation procedure for MRVs that is less affected by NRVs.

Purpose

The primary goal of this study was to tie together various lines of research in an attempt to understand how NRVs affect the estimation of several different person parameters for MRVs. Having many NRVs in a calibration sample will cause poor item parameter estimation, but the number and type of NRVs necessary to cause inaccuracy in $\hat{\theta}$, l_z, and ZU3 for MRVs has yet to be determined. A secondary goal of this study was to evaluate the distributional characteristics of the l_z index and compare it to the nonparametric ZU3 index. Studying the distributional characteristics of person fit will provide insight into how the context in which an examinee is evaluated affects not only the accuracy of $\hat{\theta}$ but also how well an examinee appears to fit the underlying test model.

Method

A sample of 10,000 examinees was drawn from an administration of the ACT math test. This original dataset was fit to the three parameter IRT model using BILOG where item parameter estimates were obtained for all 60 items. 10,000 new θ values were randomly drawn from a N(0,1) distribution, and

using monte carlo procedures a 60 item 0/1 response vector was simulated for each θ value. This 10,000 by 60 response data matrix served as an initial dataset from which NRVs could be simulated.

Actual response data from real examinees was not used here, because the number of NRVs and the type of NRVs would not have been known. By simulating response vectors we know that all the examinees (probabilistically) fit the underlying model. Real item parameters were used, so that the results reported here would be more like what would be found in a real testing situation compared to using contrived item parameters [e.g., uniformly distributed in difficulty (see Davey, Nering, & Thompson, 1997, for a discussion of this issue)]. Also, NRVs were simulated so as to represent response behaviors that might be found in a real testing situation (Meijer, 1996).

The 10,000 by 60 initial dataset was fit to BILOG where $\hat{\theta}$ values were found for each examinee using an MLE, BIW, and EAP estimation procedure. These initial $\hat{\theta}$ values served as a basis of comparison to which they were evaluated against $\hat{\theta}$ values found from datasets where NRVs had been simulated. Also, for each simulee initial l_z and ZU3 indices were determined, and the first four moments of the distributions of l_z and ZU3 were found. For each examinee three different l_z values were determined, where $\hat{\theta}$ in Equations 2 through 4 were found using either MLE, BIW, or EAP.

Simulation Procedures

As discussed above, there is a variety of response behaviors that may result in the IRT model not fitting an examinee's responses. Meijer (1996) suggests that there are two response behaviors that may be of particular interest, namely cheating and guessing. These two behaviors have always been challenging for measurement specialists and it is important to understand what influence they have on other examinees. As in the Meijer study these behaviors were simulated under several different experimental conditions. Two different levels of cheating were simulated, cheating on 20% of the most difficult items and cheating on 20% of the items with middle difficulty. In the cheating conditions responses to selected items were changed to correct, regardless of the original monte carlo generated response. Two levels of guessing were also simulated where examinees either guessed on 20% of the items or they guessed on all

the items, where item responses were changed with a 20% chance to correct response. Thus, there were a total of four response manipulation conditions.

Because the number of NRVs is important, four different proportions of NRVs were studied. The initial dataset was subjected to the four response manipulation conditions where so that either 5, 10, 15, or 20% of the examinees were manipulated to not fit the model. Thus, a total of 16 different response manipulated datasets were studied (4 response manipulation conditions × proportion of NRVs).

Evaluation

Bias charts similar to those Meijer (1996) used were constructed to study the influence that the NRVs had on MRVs. These charts were constructed not only to determine how the type and number of NRVs influenced estimation for the MRVs, but also to study how accurate $\hat{\theta}$ was using the different ability estimation procedures. For each condition studied MRVs were grouped by their θ level, where simulees with $-1.75<\theta<-1.25$ were grouped at the θ=-1.5 level, simulees with $-1.25<\theta<-0.75$ were grouped at θ=-1.0, and so forth. This was done so that there were several thousand examinees in each group so that bias statistics were less influenced by sampling error. Bias for each group of simulees was computed by the typical average signed difference (ASD) formula:

$$ASD = \frac{\sum_{j=1}^{K}\left(\hat{\theta}_j^{*} - \hat{\theta}_j\right)}{K}, \qquad (11)$$

where

j indexes the persons in an ability group (j = 1, 2, ..., K),

$\hat{\theta}_j^{*}$ represents the ability estimates found in the initial dataset, and

$\hat{\theta}_j$ represents the ability estimated in the condition with NRVs present.

The distributions of the l_z and the ZU3 indices were evaluated by the first four moments of their distributions for each condition. As with the ASD charts, the distributional characteristics were evaluated

only for the MRVs. Because we are interested in the change in the distribution of person fit for MRVs that may be caused by NRVs, ASD and root mean squared difference (RMSD) statistics were calculated for the l_z and ZU3 indices for the MRVs. The RMSD values were found for l_z, for example, by:

$$RMSD = \sqrt{\frac{\sum_{J=1}^{K}\left(l_z^* - l_z\right)^2}{K}}. \qquad (12)$$

The ASD and RMSD statistics for the person fit indices were found using the same method that was used for θ by subtracting the person fit index found in the various conditions (l_z) from the person fit index found in the initial dataset (l_z^*).

Results

Ability Estimation

In Figures 2 through 5 are ASD charts found for the various conditions studied. The ASD due to different amounts of cheating on 20% of the most difficult items is presented in Figure 2, where the different panels represent the different estimation methods. By comparing the graphs in Figure 2, it is obvious that the amount of ASD in the EAP estimates (Figure 2c) was much lower at the negative end of the θ continuum compared to when MLE and BIW were used (i.e., Figures 2a and 2b). ASD was largest (over 0.5) when MLE and BIW estimation were used and 20% of the simulees in the calibration sample were NRVs for examinees in the θ=-1.5 group. It is important to note that the bias presented here is relative bias, that is the bias between $\hat{\theta}$ values found for MRVs under different conditions. Thus, the EAP estimates, for example, contain the usual bias relative to true θ that has been found by previous researchers (e.g., McBride, 1977).

Interestingly, ASD was larger when examinees cheated on 20% of the medium difficulty items (Figure 3), than when they did so on 20% of the most difficulty items (Figure 2). Simulees on the negative end of the continuum were much more affected. For example, in Figure 3a where MLE was used

and when there were 10% NRVs ASD was close to 0.0 for all θ>-0.5, but ASD was approximately 0.4 for simulees located around θ=-1.5. The ASD was smallest when an EAP method was used, but for simulees where θ<= -1.0 ASD was approximately 0.36 (see Figure 3c). ASD in Figures 3a and 3b found when MLE and BIW were used was extremely large for simulees below θ=0.0. Even with only 10% of the simulees not fitting the model in the calibration sample, for simulees around θ=-1.5 ASD was larger than 0.30.

Figures 4 and 5 contain the ASD charts for the conditions where NRVs in the calibration sample were guessing on some or all of the items. In Figure 4 where there were different numbers of simulees guessing on all the items, the estimation procedure used appeared to have less of an influence on the ASD compared to when simulees were cheating (Figures 2 and 3). However, as with Figures 2 and 3 there was a larger amount of ASD for MRVs located at the negative end of the θ continuum. With only 5% of the simulees in the calibration sample and guessing on all the items, the ASD for MLE and BIW at θ=-1.5 was 0.25, and for the EAP procedure the value was approximately 0.20. Thus, having a small proportion of the examinees guessing on all of the items can cause rather large ASD in the estimation of θ for MRVs located at the negative end of the score continuum. In Figure 5, where the NRVs were guessing on 20% of the items there was very little ASD regardless of the estimation method and regardless of how many simulees were not fitting the model in the calibration sample. Thus, having NRVs guessing on a portion of the items on a test did not affect $\hat{\theta}$ for MRVs.

Distribution of Person Fit

Null condition. In Table 1 are the distributional characteristics of the l_z and ZU3 found from the initial (i.e., model fitting) dataset. As discussed above, the distributions of l_z and ZU3 should have an expected value of 0.0, and a standard deviation of 1.0. When an EAP estimation procedure was used the mean of the l_z distribution was 0.126, which was slightly larger compared to when MLE (0.024) or BIW (0.018) were used. However, the mean of ZU3 (-0.386) was quite different compared to what was expected. Interestingly, the standard deviation of l_z was less than expected (around 0.85 for all conditions) while the standard deviation for ZU3 was larger (1.255) than expected. Also, indices of

skewness and kurtosis where much larger for l_z compared to ZU3 (columns 3 and 4 of Table 1). For example, indicies of skewness for l_z ranged from -0.263 to -0.402, while for ZU3 the index of skewness was -0.103.

The l_z Index. In Tables 2 and 3 are the distributional characteristics, along with the ASD and RMSD, for l_z found in the various cheating conditions studied. Comparing the distributional characteristics in Table 2 to Table 1, having NRVs where cheating was simulated in the calibration sample clearly affected the distribution of l_z for MRVs. For example, in the case where 10% of the sample contained response vectors that were cheating in 20% of the most difficult items, the mean of l_z was consistently above 0.2 regardless of the estimation procedure used. While the SD values presented in Table 2 tended to be close those values found in the null condition (all around 0.85), there were cases where the SD was smaller than what was expected. For example, when 20% of the simulees were cheating on the most difficult items and when EAP was used the SD of l_z was 0.783. Comparing the top half of Table 2 to the bottom half, the distribution of l_z appeared to be most affected when simulees were cheating on the most difficult items. However, when 20% of the simulees were cheating on the most difficult items, the means of l_z when MLE and BIW were used (0.046 and -0.001, respectively) were closer to what was expected compared to when simulees where cheating on the medium difficulty items (-0.149 and -0.168, respectively). Also, the SD values for l_z tended to be closer to what was expected when simulees were cheating on the most difficult items compared to the medium difficulty items when BIW was used. For example, in the conditions where 20% of the simulees were cheating the SDs of l_z when BIW was used were 0.967 and 0.905 when cheating occurred on the most difficult items and the medium difficulty items, respectively.

The indices of skewness and kurtosis in Table 2 suggest that, in general, the distribution of l_z followed a normal distribution when cheating response vectors were present, except when BIW was used. The kurtosis, in particular, suggests that the BIW procedure results in a sometimes dramatic level of kurtosis (e.g., over 26 when 20% cheating on the most difficult items). The distribution of l_z where 20% of the simulees were cheating on the most difficult items and when BIW was used is presented in Figure 6 (Also plotted in this figure is a normal curve to serve as a reference). Notice that ###. The ASD and

RMSD values presented in Table 2 also demonstrated that having NRVs in the calibration sample tended to affect the distribution of l_z for MRVs compared to that found in the null condition. As with the distributional characteristics larger values of ASD and RMSD were found when the cheating occurred on the most difficult items. One interesting finding occurred when comparing ASD values found in the 15% cheating conditions to the 20% cheating conditions. In these conditions the ASD values were much closer to 0.0 when more examinees were cheating, for example, when EAP was used ASD changed from -0.206 to -0.098 as the number of cheating NRVs increased from 15% to 20%.

Table 3 contains the distributional characteristics for l_z found in the guessing conditions. After a comparison of Tables 2 and 3 when NRVs were cheating rather than guessing the distribution of l_z was obviously much more affected. Overall, the mean and SD values of l_z were much closer to 0.0 and 1.0, respectivley, especially when NRVs were guessing on all the items. In the conditions where NRVs were guessing on only 20% of the items the mean values tended to be larger compared to when NRVs were guessing on all the items. For example, the mean varied from 0.168 to 0.274 when 15% NRVs were guessing on 20% of the items, and 0.101 to 0.173 when NRVs were guessing on all the items. The indices of skewness and kurtosis followed a similar pattern compared to the cheating conditions, where values were largest when BIW was used. Interestingly, |ASD| values were consistently larger when NRVs were guessing on 20% of the items, while the RMSD values were less systematic.

The ZU3 Index. The distributional characteristics of the ZU3 index are presented in Table 4. Compared to the null condition (Table 1), the ZU3 index continued to have relatively larger negative means and SD values than expected in the experimental conditions. The mean of ZU3 was consistently above -0.40 when cheating occurred on the most difficult items, but was much closer to the -0.386 value found in the null condition (Table 1) when cheating occurred on the medium difficulty items. As with l_z, the mean of ZU3 was quite different from the null condition when examinees where guessing on all items compared to when they were guessing on only 20% of the items. For example, when there were 15% NRVs the mean changed from 0.170 to -0.324 when simulees were guessing on all and 20% of the items, respectively. The SD values were slightly more consistent compared to the mean values, and ranged from 1.143 to 1.305.

The indices of skewness and kurtosis presented in Table 4 demonstrated that the ZU3 index tended to follow a more normal distribution compared to the l_z index (Tables 2 and 3). The skewness, in particular, was very small with values ranging from -0.01 to -0.201. The largest values of kurtosis were found when NRVs were guessing on all items, with the largest value being -0.598 in the condition where there were 20% NRVs. In most conditions studied, there was not a systematic change in ZU3 for MRVs. This can be seen in the ASD values in Table 4, where they tended to be less than 0.1 (in absolute value), except for the conditions where simulees were guessing on all the items. In these conditions ASD values were much larger (e.g., -0.544 with 15% NRVs) as were the RMSD values (e.g., 2.542 observed in the same condition).

Discussion

This study extends the findings presented by Meijer and Nering (in press) and Reise (1995), and demonstrated that particular ability estimation procedures may be more robust to nonmodel-fitting responses compared to other estimation procedures. In this study (unlike what Meijer & Nering and Reise found), the EAP procedure tended to result in less bias in the ability estimate; however, the distribution of l_z tended to more closely approximate a standard distribution (i.e., mean=0.0 and SD=1.0) when the BIW procedure was used compared to the MLE and EAP procedures. Unfortunately, the indicies of skewness and kurtosis presented in Tables 2 and 3 show that the l_z index did not at all follow a normal distribution when BIW was used. Thus, the relationship between the distribution of person fit and the accuracy in $\hat{\theta}$ for MRVs was not systematic under the various conditions studied here.

For the l_z and the ZU3 indices there did not appear to be a large difference in the distributions when examinees were either cheating or guessing. In these respective conditions the distributions were more affected when cheating occurred on the most difficult items, or when NRVs were guessing on 20% of the items. The results for cheating conditions are not surprising. One possible explanation for why the distributions may have been less affected by guessing on all items may be because in this condition there was a large amount of estimation error (comparing Figures 4 and 5), causing the distribution of person fit

to not reflect what was really happening. Certainly, additional work is needed to further evaluate this issue.

In a real testing situation a dataset will not contain examinees that are either guessing or cheating. In this study these types of nonmodel-fitting behaviors were studied under separate conditions, and results from a real testing situation will more than likely be more complicated. Additional work is needed where different types of NRVs are blended together in a calibration sample to determine how accurately person parameters are estimated for MRVs in a more realistic testing situation. Also in this study very large sample sizes were used, and in a real testing situation error in estimated person parameters may be exacerbated by small sample calibration.

Previous work has not closely examined the distribution of the ZU3 index, and the results of this study were promising. Although the distribution of ZU3 did not typically have mean and SD values close to what was expected, this index appeared to be much less affected by NRVs compared to l_z. One reason may be than in order to calculate l_z it was necessary (at least in this study) to estimate three parameters for each item. However, for ZU3 only the π_g values needed to be calculated (Equations 6 through 10) for each item. Thus, the inaccuracy in l_z may be the result of accumulation of estimation error across the various item parameters.

The findings in this study suggest that researchers should seriously consider the issue of whether there may be NRVs in a calibration sample. The results presented here demonstrate that MRVs may have poorly estimated ability levels, and the extent to which they appear to fit the underlying model may not be accurately estimated, when NRVs are present. One possible solution might be to run initial calibrations to identify possible NRVs, remove these examinees, and recalibrate the dataset. This two-stage calibration process may lead to more accurate item parameters, and thus more accurate person parameters.

References

Birenbaum, M. (1985). Comparing the effectiveness of several IRT based appropriateness measures in detecting unusual response patterns. *Educational and Psychological Measurement, 45,* 523-534.

Birenbaum, M. (1986). Effect of dissimulation motivation and anxiety on response pattern appropriatness measures. *Applied Psychological Measurement, 10,* 167-174.

Davey, T., Nering, M. L., & Thompson, T. (1997). *Realistic simulation of item response data.* Presented at the annual meeting of the National Council of Measurement in Education, Chicago, IL.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11,* 59-79.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15,* 171-191.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indicies. *British Journal of Mathematical and Statistical Psychology, 38,* 67-86.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4,* 269-290.

Li, M. F., & Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement, 21,* 215-231.

McBride, J. R. (1977). Some properties of a Bayesian adaptive ability testing strategy. *Applied Psychological Measurement, 1,* 121-140.

Meijer, R. R. (1996). The influence of the presence of deviant item score patterns on the power of a person-fit statistic. *Applied Psychological Measurement, 20,* 141-154.

Meijer, R. R. (1997). Person fit and criterion related validity: An extension of the Schmitt, Cortina, and Whitney study. *Applied Psychological Measurement, 21,* 99-113.

Meijer, R. R., & Nering, M. L. (in press). Trait level estimation for nonfitting-response vectors. *Applied Psychological Measurement.*

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education, 8,* 261-272.

Mislevy, R. J., & Bock, R. D. (1982). Biweight estimators of latent ability. *Educational and Psychological Measurement, 42,* 725-737.

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* New York: De Gruyter.

Mokken, R. J., & Lewis, C. (1983). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6,* 417-430.

Nering, M. L. (1997). The distribution of person fit in the computerized adaptive testing environment. *Applied Psychological Measurement, 21,* 115-127.

Nering, M. L. (1996). *The effects of person misfit in computerized adaptive testing.* Unpublished doctoral dissertation. University of Minnesota.

Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19,* 121-129.

Nering, M. L., & Meijer, R. R. (in press).

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 19,* 213-229.

Reise, S. P., & Due, A. M. (1991). Test characteristics and their influence on the detection of aberrant response patterns. *Applied Psychological Measurement, 15,* 217-226.

Reise S. P., & Waller, N. G. (1993). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14,* 45-58.

Schmitt, N., Cortina, J. M., & Whitney, D. J. (1993). Appropriateness fit and criterion-related validity. *Applied Psychological Measurement, 17,* 143-150.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49,* 95-110.

Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item characteristic curve models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing.* New York: Academic Press.

van der Flier, H. (1982). Deviant response patterns and comparability of test score. *Journal of Cross-Cultural Psychology, 13,* 267-298.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 97-116.

Zickar & Drasgow, (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20,* 71-87

Figure 1
Schematic of Research Questions
(where J is the number of nonmodel-fitting simulees)

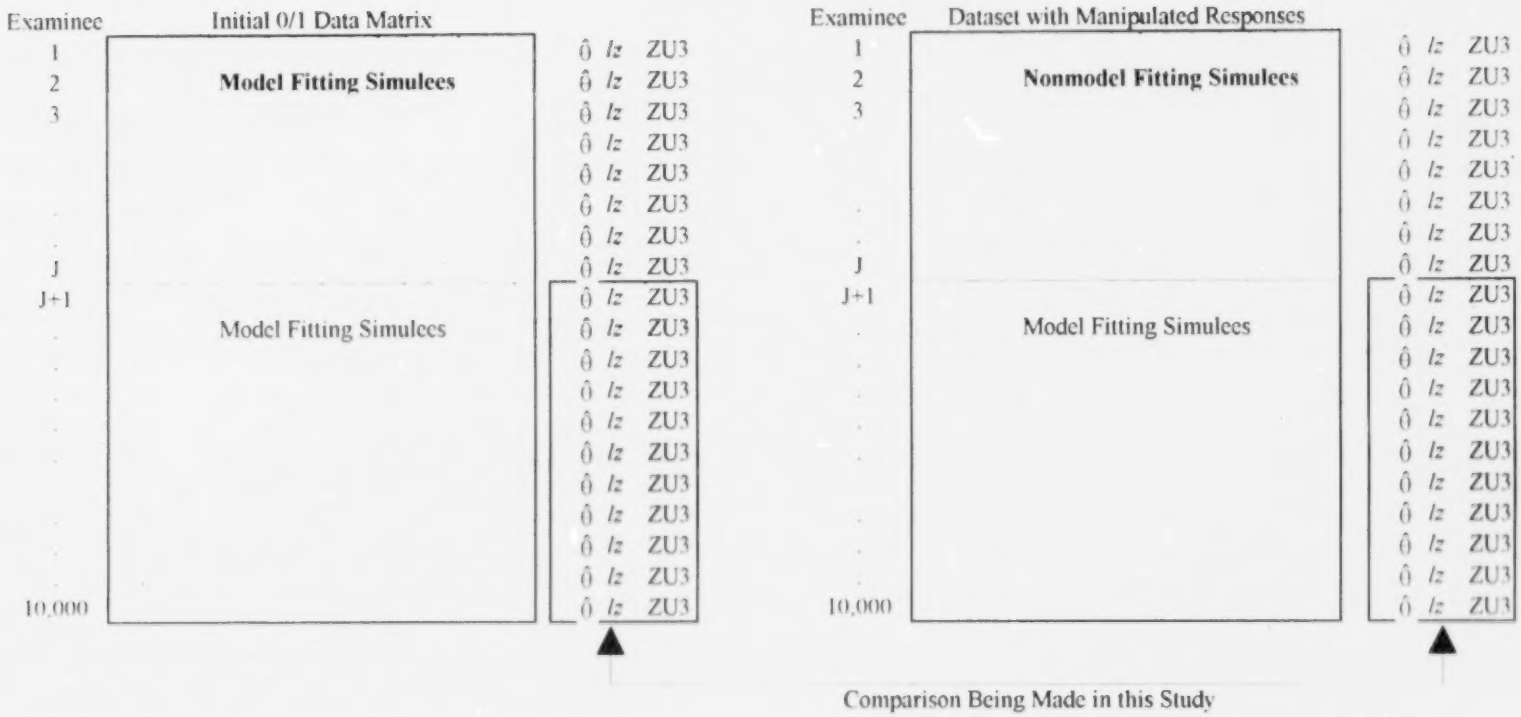
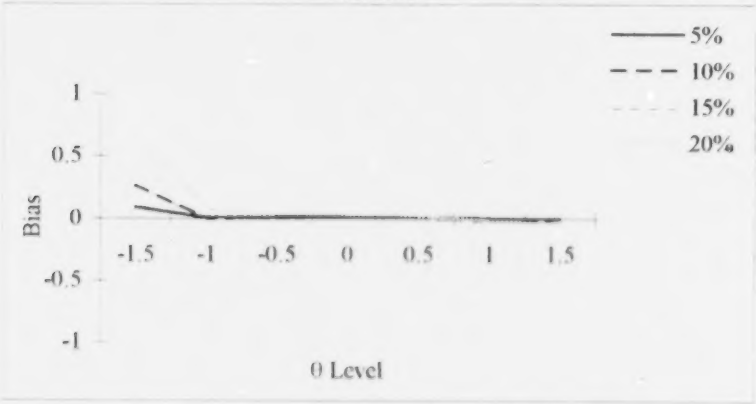
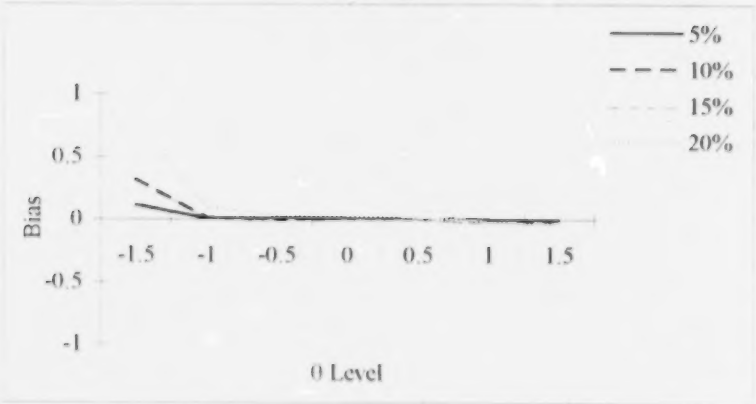


Figure 2
Bias Due to Different Amounts of Cheating on 20% of the Most Difficult Items

a. MLE



b. Biweight



c. EAP

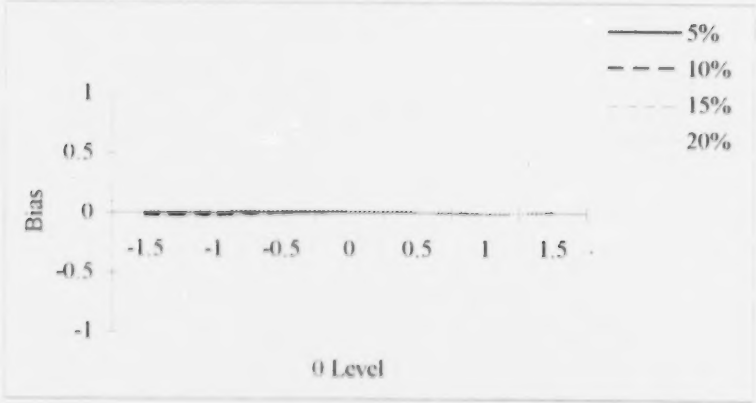
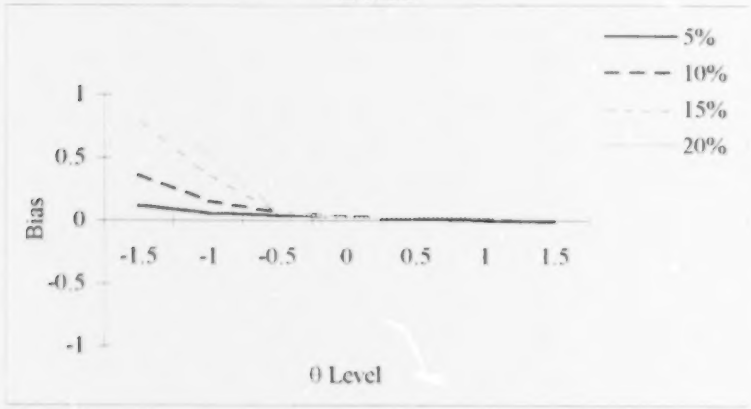
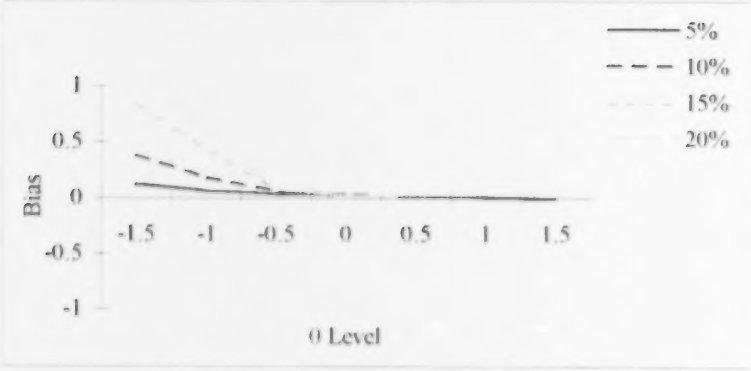


Figure 3
Bias Due to Different Amounts of Cheating on 20% of the Medium Difficult Items

a. MLE



b. Biweight



c. EAP

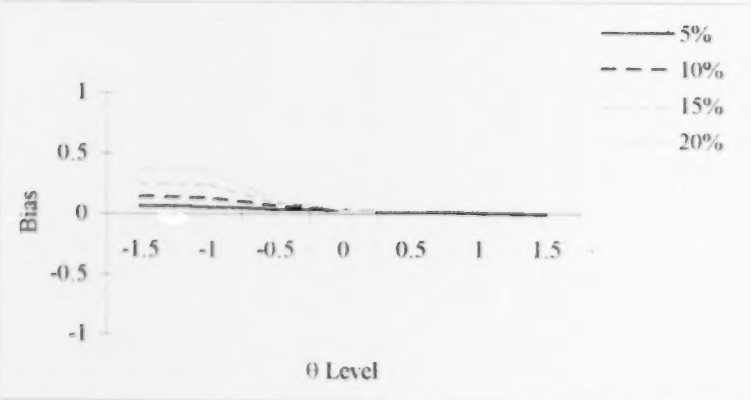
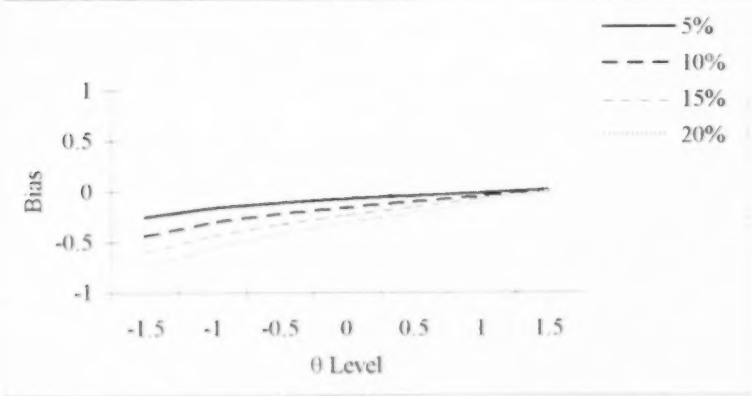
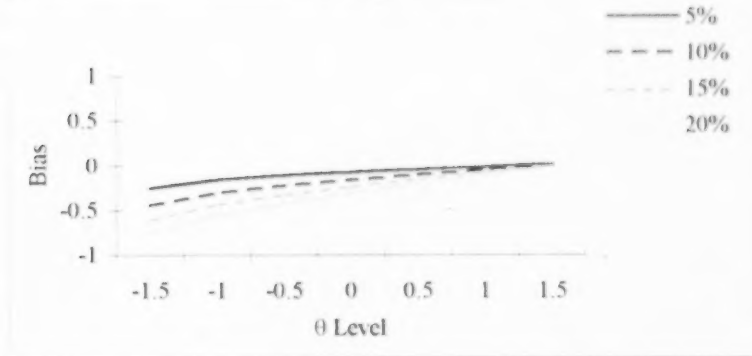


Figure 4
Bias Due to Different Amounts of Guessing on All Items

a. MLE



b. BIW



c. EAP

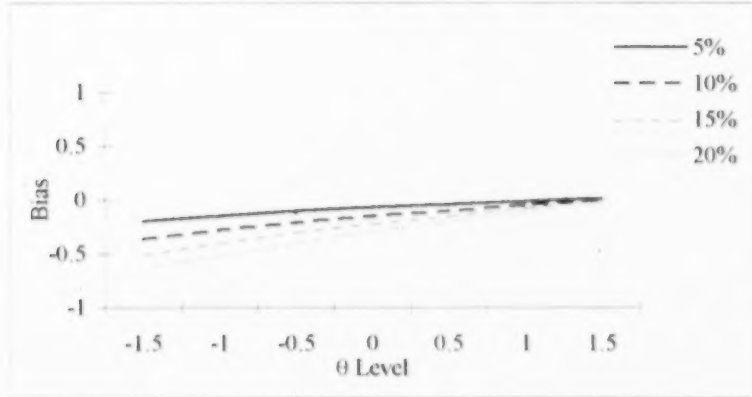


Figure 5
Bias Due to Different Amounts of Guessing on 20% of Items

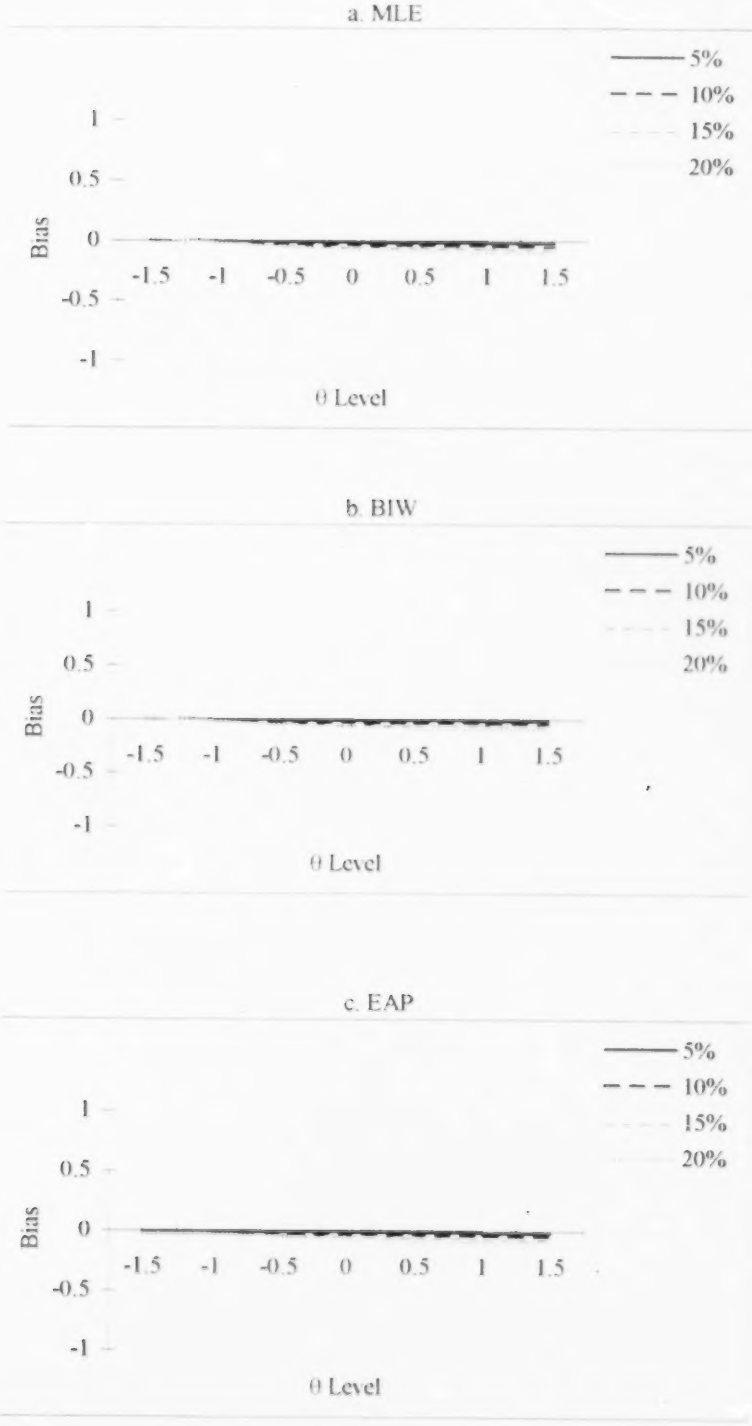


Table 1
Distributional Characteristics of l_z and ZU3 with all Model Fitting Response Vectors

Person Fit Index and Estimation Method	Mean	SD	Skewness	Kurtosis
l_z				
MLE	0.024	0.838	-0.263	0.551
BIW	0.018	0.885	-0.402	0.983
EAP	0.126	0.851	-0.368	0.499
ZU3	-0.386	1.255	-0.103	-0.198

Table 2
Distributional Characteristic and Change in l_z from Original Sample

% NRVs & Estimation Method	Mean	SD	Skewness	Kurtosis	ASD	RMSD
Cheating on 20% most difficult						
5%						
MLE	0.187	0.810	-0.146	0.506	-0.169	0.254
BIW	0.179	0.865	-0.430	1.676	-0.169	0.264
EAP	0.295	0.819	-0.276	0.493	-0.176	0.232
10%						
MLE	0.217	0.811	-0.071	0.409	-0.203	0.383
BIW	0.203	0.882	-0.565	2.879	-0.196	0.414
EAP	0.349	0.813	-0.219	0.468	-0.230	0.352
15%						
MLE	0.181	0.804	-0.079	0.446	-0.171	0.438
BIW	0.150	0.908	-1.015	6.828	-0.149	0.524
EAP	0.319	0.798	-0.240	0.563	-0.206	0.406
20%						
MLE	0.046	0.793	-0.204	0.471	-0.040	0.535
BIW	-0.001	0.967	-2.426	26.288	-0.002	0.706
EAP	0.208	0.783	-0.372	0.650	-0.098	0.467
Cheating on 20% medium difficult						
5%						
MLE	0.069	0.829	-0.274	0.561	-0.051	0.136
BIW	0.064	0.876	-0.413	0.970	-0.054	0.148
EAP	0.171	0.844	-0.383	0.504	-0.051	0.124
10%						
MLE	0.029	0.830	-0.311	0.539	-0.014	0.235
BIW	0.020	0.879	-0.438	0.904	-0.013	0.257
EAP	0.139	0.851	-0.407	0.470	-0.021	0.218
15%						
MLE	-0.056	0.834	-0.356	0.500	0.065	0.351
BIW	-0.073	0.884	-0.466	0.870	0.074	0.380
EAP	0.063	0.871	-0.404	0.365	0.050	0.331
20%						
MLE	-0.149	0.854	-0.418	0.432	0.154	0.458
BIW	-0.168	0.905	-0.519	0.892	0.165	0.485
EAP	-0.043	0.902	-0.412	0.290	0.154	0.453

Figure 6

Table 3
Distributional Characteristic and Change in l_z from Original Sample

% NRVs & Estimation Method	Mean	SD	Skewness	Kurtosis	ASD	RMSD
Guessing on all items						
5%						
MLE	0.084	0.835	-0.279	0.562	-0.063	0.129
BIW	0.076	0.894	-0.590	2.426	-0.062	0.152
EAP	0.162	0.846	-0.367	0.516	-0.040	0.073
10%						
MLE	0.102	0.838	-0.267	0.560	-0.081	0.162
BIW	0.090	0.915	-0.886	5.901	-0.076	0.227
EAP	0.170	0.848	-0.345	0.513	-0.047	0.107
15%						
MLE	0.117	0.838	-0.248	0.565	-0.097	0.186
BIW	0.101	0.936	-1.305	12.011	-0.088	0.290
EAP	0.173	0.845	-0.327	0.520	-0.052	0.139
20%						
MLE	0.133	0.842	-0.239	0.562	-0.114	0.211
BIW	0.112	0.965	-2.019	26.141	-0.100	0.354
EAP	0.177	0.845	-0.316	0.524	-0.056	0.168
Guessing on 20% of the items						
5%						
MLE	0.097	0.825	-0.270	0.585	-0.076	0.109
BIW	0.093	0.872	-0.427	1.074	-0.079	0.112
EAP	0.193	0.839	-0.384	0.546	-0.071	0.099
10%						
MLE	0.133	0.818	-0.253	0.601	-0.112	0.148
BIW	0.132	0.863	-0.394	0.966	-0.118	0.158
EAP	0.235	0.831	-0.379	0.571	-0.112	0.144
15%						
MLE	0.168	0.810	-0.235	0.631	-0.148	0.188
BIW	0.170	0.851	-0.361	0.891	-0.157	0.208
EAP	0.274	0.821	-0.375	0.615	-0.153	0.189
20%						
MLE	0.204	0.809	-0.232	0.622	-0.185	0.237
BIW	0.209	0.848	-0.343	0.749	-0.197	0.262
EAP	0.311	0.817	-0.375	0.642	-0.189	0.231

Table 4
Distributional Characteristic and Change in ZU3 from Original Sample

Type of Misfit & % NRVs	Mean	SD	Skewness	Kurtosis	ASD	RMSD
Cheating on 20% most difficult						
5%	-0.421	1.231	-0.025	-0.244	0.036	0.159
10%	-0.444	1.217	0.026	-0.242	0.058	0.281
15%	-0.440	1.190	0.049	-0.190	0.060	0.370
20%	-0.406	1.143	0.007	-0.140	0.033	0.445
Cheating on 20% medium difficult						
5%	-0.386	1.243	-0.087	-0.205	0.002	0.082
10%	-0.382	1.238	-0.110	-0.172	-0.003	0.179
15%	-0.360	1.234	-0.142	-0.081	-0.020	0.300
20%	-0.318	1.240	-0.201	0.031	-0.055	0.446
Guessing on all items						
5%	-0.305	1.237	0.009	-0.373	-0.078	0.180
10%	-0.232	1.246	0.062	-0.488	-0.150	0.338
15%	0.170	1.268	-0.063	-0.555	-0.544	2.542
20%	0.128	1.305	-0.055	-0.598	-0.507	2.553
Guessing on 20% of the items						
5%	-0.367	1.245	-0.069	-0.245	-0.017	0.041
10%	-0.347	1.242	-0.038	-0.295	-0.035	0.081
15%	-0.324	1.240	-0.028	-0.330	-0.051	0.120
20%	-0.311	1.243	-0.010	-0.368	-0.068	0.158

Acknowledgments

The authors wish to expresses their appreciation to Amy Hendrixson for her careful editing and to Tim Burden, Michael Finger, Rob R. Meijer, Tim Miller, and Qing Yi for their helpful comments on an earlier version of the paper.

Author's Address

Send requests for reprints or futher information to Michael L. Nering, ACT, 2201 North Dodge St., PO Box 168, Iowa City, IA 52243-0168. Email: nering@act.org.